

METHOD, APPARATUS AND ARTICLE FOR DATA REDUCTION

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to methods, apparatus and articles used for the analysis of data. More particularly, it relates to methods, apparatus and articles used to reduce large amounts of data to much smaller forms. More particularly still, it relates to methods, apparatus and articles which may be used to reduce the data to a form conducive to efficient analysis of the data.

2. Prior Art

There are many instances in which a great deal of data is produced, and it is a lengthy and costly process to analyze the data. By way of example only, and not by way of limitation, in the broadcast industry, it is desirable to monitor the programs broadcast by stations to determine whether specific content, such as for example music or songs, have been broadcast.

An example of how this may be accomplished is disclosed in United States Patent No. 5,437,050, entitled Method and Apparatus for Recognizing Broadcast Information Using Multi-Frequency Magnitude Detection, issued to Lamb et al. and assigned to the same assignee as that of the present invention.

As noted in this patent, a wide variety of copyrighted recordings and commercial messages are transmitted by broadcast stations. Copyrighted works such as moving

pictures, television programs, and phonographic recordings attract audiences for broadcast stations, and the aforementioned commercial messages, when sent to the audiences, provide revenue for the broadcast stations.

5

There is an interest among various unions, guilds, performance rights societies, copyright owners, and advertising communities in knowing the type and frequency of information being broadcast. Owners of copyrighted works, for example, may be paid a royalty rate by broadcast stations depending on how often their copyrighted work is broadcast. Similarly, commercial message owners such as advertisers, who pay broadcast stations for air time, have an interest in knowing how often their commercial messages are broadcast.

It is known in the art that commercial radio and television broadcast stations are regularly monitored to determine the number of times certain information is broadcast. Various monitoring systems have been proposed in the prior art. In manual systems, which entail either real-time listening or delayed listening via video or audio tapes, people are hired to listen to broadcast information and report on the information they hear. Manual systems although simple, are expensive, lack reliability, and are very often highly inaccurate.

Electronic monitoring methodologies offer advantages over manual systems such as lower operating costs and reliability. One type of electronic monitoring methodology requires insertion of specific codes into broadcast information before the information is

transmitted. The electronic monitoring system can then recognize a song, for example, by matching the received code with a code in a reference library. Such systems suffer from both technical and legal difficulties. For
5 example, such a coding technique requires circuitry, which is expensive to design and assemble and which must be placed at each transmitting and receiving station. Legal difficulties stem from the adverse position of government regulatory agencies toward the alteration of
10 broadcast signals without widespread acceptance thereof by those in the broadcast industry.

A second type of electronic monitoring methodology requires pre-specification of broadcast information into
15 a reference library of the electronic monitoring system before the information can be recognized. A variety of pre-specification methodologies have been proposed in the prior art. The methodologies vary in speed, complexity, and accuracy. Methodologies which provide accuracy are
20 likely to be slow and complex, and methodologies which provide speed are likely to be inaccurate.

The apparatus and method described in the above mentioned United States Patent No. 5,437,050 has met with
25 commercial success and has in large part met the needs of many segments of the broadcast industry. This approach is based on the discovery that the broadcast information on which recognition is based lies in the narrow frequency bands associated with the semitones of the
30 music scale, rather than in the continuum of audio frequencies or in other sets of discrete frequency bands. It is also based on the principle that the set of

semitones that have energies above a threshold amount at each instant provide sufficient information for recognition, and that it is not necessary to use the absolute energies of all frequencies for recognition.

5

Thus, United States Patent No. 5,437,050 provides apparatus and a method of recognizing broadcast information, including the steps of receiving broadcast information, the broadcast information being in analog
10 form and varying with time; converting the broadcast information into a frequency representation of the broadcast information; dividing the frequency representation into a plurality of separate frequency bands (generally 48 bands over four octaves); determining
15 a magnitude of each separate frequency band of the digital sample; and storing the magnitudes. The method of recognizing broadcast information also includes the steps of performing a significance determination a plurality of times, the significance determination including the steps
20 of generating a magnitude of each separate frequency band, using a predetermined number of previously stored magnitudes for each respective frequency band; storing the magnitudes; and determining a significance value, using a predetermined number of previously stored
25 magnitudes for each respective frequency band. The method of recognizing broadcast information further includes the steps of comparing the significance value to the most recently generated magnitude of each separate frequency bands generating a data array, the data array having a
30 number of elements equal to the number of separate frequency bands, the values of the elements being either binary 1 or binary 0 depending on the results of the

comparison; reading a reference data array, the reference data array having been generated from reference information; comparing the data array to the reference data array; and determining, based on the comparison,
5 whether the broadcast information is the same as the reference information.

United States Patent No. 5,437,050 also provides a digital recording method in conjunction with the
10 monitoring system to achieve recognition of broadcast information pre-specified to the monitoring system. The digital recording method can also achieve recognition of broadcast information not previously known to the monitoring system, while preserving a complete record of
15 the entire broadcast period which can be used for further reconciliation and verification of the broadcast information.

More specifically, the method of recording broadcast
20 information, includes the steps of receiving a set of broadcast information; recording the set of broadcast information in a compressed, digital form; generating a representation of the set of broadcast information; comparing the representation to a file of
25 representations; making a determination, based on the comparison, of whether the representation corresponds to any representations in the file; upon a determination that the representation corresponds to a representation in the file, recording the broadcast time, duration, and
30 identification of the set of broadcast information that corresponds to the representation; upon a determination that the representation does not correspond to any

representations in the file, performing the following steps: (a) performing a screening operation on the representation in order to discern whether the representation should be discarded; (b) upon a
5 determination that the representation should not be discarded, performing the following steps: (c) playing the recorded set of broadcast information which corresponds to the set of broadcast information from which the representation was generated in the presence of
10 a human operator; and (d) making a determination, based on the playing of the recorded set of broadcast information, of whether the representation should be added to the file of representations and whether a recording should be made of the broadcast time, duration,
15 and identification of the set of broadcast information that corresponds to the representation.

As noted above, while the technology described in United States Patent No. 5,437,050 has been widely used, over
20 the years it has become apparent that it has some limitations. While working well in its intended application, attempts to apply it to other applications have met with varying degrees of success. The technology is not extremely effective at short-term matching; that
25 is determining whether a match exists between the data produced by a short segment of source material and a previously stored reference. In general, relatively few bits are set; on average, only 3 bits in a 48 bit frame. Often, the same bits are set for many frames in a row.
30 When a strong melody is not present, continuity may be weak. Further, in the presence of a strong melody, it is possible that only one bit will be set for many frames.

Slightly different results may be produced, in terms of the specific bits set for the same source material, when different amounts of audio compression are used. Further, different bits may be set when transients are suppressed.

SUMMARY OF THE INVENTION

It is an object of the invention to provide an apparatus and method for processing a signal so as to represent the characteristics of that signal in a compact form.

It is a further object of the invention to provide an apparatus and method for reducing data to a form that supports efficient processing of the data.

It is a further object of the invention to provide an apparatus and method for determining whether the signal contains particular content, and to do so in a computationally efficient manner.

It is yet another object of the invention to provide an article of manufacture containing a computer program which causes a computer to achieve the above mentioned objects.

The present invention may be thought of as a lossy data reduction technique. A series of input frames is each comprised of a set of N scalar values which may represent, for example, amplitude, magnitude or intensity of some characteristic of an original signal S. The

nature of signal S, the choice of N characteristics, and the decomposition of signal S into input frames may take many different forms, and may be considered as independent of the invention. The input frames have a
5 sampling rate S_r . The invention produces a series of output frames with a sampling rate of that of the input frames divided by W (that is S_r/W), where W is an averaging window, represented by a whole number greater than one, and wherein each output frame comprises N bits.

10 The invention transforms the input frames into a fewer number of output frames using the following method. Each input frame is analyzed, and the top X (where X is less than N) of the N characteristic values are identified. Values in the input frames not in the top X are set to
15 zero. Subsequent input frames are processed in the same manner until W frames have been so processed, with their top values identified and values other than their top values set to zero. The respective processed values for each N in the W processed input frames are averaged,
20 producing N average values. These average N values are analyzed, and the top Y values (where Y is less than X) are identified. An output frame is produced, including one bit for each N, wherein the bit is set to one if the value of the particular average was in the top Y, and the
25 bit is set to zero if it was not in the top Y. Processing then continues for the next W input frames to produce the next output frame.

The resulting output has the following useful characteristics. First, it is much smaller amount of
30 data than the input. Specifically, the resulting size of

the data output data is $1/(Nbits*W)$ times the size of the input data, where Nbits is the number of bits used to represent each of the scalar N values. Typical values of Nbits and W are 32 and 5 respectively, yielding a 1:160
5 reduction.

In addition, the bits set to 1 in the output data tend to represent the most salient characteristics of the input signal during the time period covered by each output frame. Further, a constant number of bits, Y, is set in
10 each frame, making the resulting data and its properties amenable to straightforward analysis. The resulting output is more robust, in the sense that it is less impacted by noise, transients, and distortions than a conventional averaging technique. The output data can be
15 used in many of the applications that would require the input data. These applications include signal comparison, feature detection, pattern recognition, anomaly detection, trend analysis, etc. A significant increase in processing speed is provided, due to the
20 reduction in the amount of data that must be processed, and the fact that bit comparison operations can be used to process the data.

The invention has commercial application in the field of audio recognition, where the characteristics are
25 amplitude measurements for semitones as determined by a smoothing sine/cosine filter bank, as described in the above referenced United States patent No. 5,437,050. In this application, $N=48$, $Nbits=32$, $Sr=50Hz$, $W=5$, $X=12$ and $Y=8$. The resulting system, when compared to the system
30 described in the patent, provides a five fold reduction

in the volume of data that must be transmitted and processed, and a ten to twenty fold reduction in recognition processing time, with no loss in recognition accuracy. In this application, the invention operates on
5 the principle that it is better to produce fewer frames having more information per frame. Thus, the frequencies having the largest magnitudes are processed, and others are suppressed.

In the broadcast recognition application, the method and
10 apparatus divide the signal into a series of frames; for each frame, divide a spectrum of the signal into a series of frequency segments; determine which of a number of frequency segments of the series of frequency segments have largest amplitudes; set a value of zero for all of
15 the frequency segments other than the number having the largest amplitudes; set a value representative of amplitude for the frequency segments having the largest amplitudes, average respective values, for a series of frames, to produce a series of average values; select a
20 number of the average values which are largest average values; and produce the digital representation by setting bits to a first binary value for the selected number of the average values, and to a second binary value for all other average values. The averaging of respective values,
25 for a series of frames, to produce a series of average values includes averaging the values of zero.

The number of frequency segments of the series of frequency segments having largest amplitudes in the
30 spectrum of the frame may be a predetermined, fixed number. The number of the average values having largest

average values that are selected may also be a predetermined, fixed number. Preferably the first binary value is one and the second binary value is zero.

5 Determining which of a number of frequency segments of the series of frequency segments have largest amplitudes in the spectrum of the frame comprises performing a Fourier transform on the signal. Preferably, a Discrete Fourier Transform is used.

10 The method and apparatus perform further processing by comparing the digital representation to a set of predefined digital representations corresponding to known content; and using results of the comparison to determine whether the signal contains the known content. A
15 reference library of digital reference representations of known content may be provided.

In the broadcast application the signal may be an audio signal representative of music or of a song.

20 The invention is also directed to a method for determining likelihood of a match between a first set of data having Y of N bits set equal to a first binary value and a remainder of the bits set equal to a second binary
25 value, and a second set of data also having Y of N bits set equal to a first binary value and a remainder of the bits set equal to a second binary value. This method comprises determining the general probabilities of Y of N bits in the first set of data and in the second set of
30 data being the same; and heuristically processing the probabilities to produce a series of match values based on the number of respective bits in the first set of data

and in the second set of data that are identical. The heuristic processing may comprise assigning a match of n out of Y values a value of 1; normalizing remaining values to the value of 1 to produce resulting numbers; 5 multiplying the resulting numbers by a constant to produce multiplied numbers; and subtracting the multiplied numbers from 1 to produce the match values. It may further comprise setting match values greater than a predetermined value to values substantially equal to 1.

10

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing aspects and other features of the present invention are explained in the following description, 15 taken in connection with the accompanying drawings, wherein:

Fig. 1 is a general, high level block diagram showing the use of the invention in data processing.

Fig. 2A, Fig. 2B and Fig. 2C are tables of data which 20 illustrate the manner in which data reduction is performed, in accordance with the invention.

Fig. 3 is a block diagram of a representation of the manner in which data may be processed, in accordance with the flow chart of Fig. 4.

Fig. 4 is a flow chart of the method of processing data 25 in accordance with the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to Fig. 1, the present invention may be used in applications having a source of data 10. This source may be comprised of data acquisition components of a broadcast monitoring system of the type described in United States Patent No. 5,437,050 (which is incorporated herein in its entirety), but is not in any way so limited. Although the present invention will be described principally with reference to this application, it should be understood that the present invention can be used in many other applications. For example, applications of the present invention include data reduction for the analysis and processing of any time-varying signal that can be decomposed into a finite number of characteristics. These applications include audio analysis where the characteristics are the output frequencies as determined by a discrete Fourier transform or a fast Fourier transform (as in the above mentioned application). Some of these applications include determining the characteristics of sound; whether is it soft or loud, music or voice, tonal characteristics such as the key in which music is played, or its tempo. Other applications include video analysis where the characteristics represent the intensities of certain spectral components, video analysis where the characteristics are signal intensity levels at certain screen locations, web site usage analysis where the characteristics are hit counts for certain pages, and traffic analysis where the characteristics are traffic volume measurements at certain intersections. The present invention may also be used for purposes of data reduction for general clustering analysis.

In Fig. 1, the data provided by data source 10 is processed by a data reduction block 20, in accordance with the invention. After processing by data reduction block 20, the reduced data is fed to a data processing block 30, where processing is performed to produce a desired result. For example, various data modeling techniques may be used. In the case of a broadcast content recognition system, the data may be processed in the manner similar, but not necessary identical, to the manner disclosed in United States Patent No. 5,437,050, which again is used merely by way of example, and not by way of limitation. Various modeling techniques may be used during the process of analyzing the reduced data, including a "scoring algorithm" as described below.

Fig. 2A. Fig. 2B and Fig. 2C illustrate the manner in which data is reduced in accordance with the invention. In Fig. 2A the amplitudes of five successive frames of data from a data source are illustrated in columns with headings Frame 1, Frame 2, Frame 3, Frame 4 and Frame 5. Each frame has six distinct values, as represented by the six rows of data. In actual practice, there may be many more rows for each frame, and many more frames could be processed simultaneously. For example, if the invention is used in the context of broadcast signal content recognition in accordance with United States Patent No. 5,437,050, then each frame would have 48 values, and thus there would be 48 rows of data. Each value would be representative of the actual amplitude of a particular frequency within the four selected octaves of the musical scale as provided by the data source.

For each frame of Fig. 2A, the top X values are selected. In Fig. 2A, X=2. These values are represented by bold faced type.

5 Referring to Fig. 2B, all values other than those selected in Fig. 2A are set equal to zero. The selected values are not changed, and remain as such in the matrix.

Referring to Fig. 2C, in a sixth column, the average value for each row is calculated, with the values set to zero averaged in as such. Then the Y largest average
10 value are selected. In this example Y=3, so that the largest three values are shown in bold-faced type. An Output is produced (column at the right) wherein the bit value for the selected Y largest values is set equal to binary 1. The remainder of the bit values are set equal
15 to binary 0. The data reduction has been completed, with Y of the original values set to binary 1. Thus, as discussed in more detail below, the number of bits having a binary 1 value in the output of the data reduction process is constant, regardless of the precise nature of
20 the data provided by the data source 10 (Fig. 1).

As noted above, source data may be derived from various sources. In the broadcast industry, source data may be program content. Reference is made the detailed description of the operation of Fig. 1 and Fig. 2 in
25 United States Patent No. 5,437,050, which, is incorporated by reference herein. By way of example only, the outputs produced by apparatus such as the forty eight notch filters of Fig. 2 of United States Patent No. 5,437,050 may be processed as illustrated in Fig. 3
30 herein and in accordance with method steps of Fig. 4 herein. Although implemented in software, it will be

recognized by those skilled in the art that all functions illustrated therein may also be performed by an appropriately designed hardware, although generally at much greater cost. The invention may be implemented in
5 software, using any general purpose programming language.

Referring to Fig. 3 and to Fig. 4 herein, the values produce by the forty eight notch filters 130a to 130i (Fig. 2 of United States Patent No. 5,437,050) for successive frames (steps 80 and 82 of Fig. 4 herein) are
10 stored in a series of respective registers R1 to R48. Just as there are forty eight notch filters, there are forty eight registers. The values stored in these registers are inspected by an arithmetic processing routine 52, and by successive comparison, or other
15 techniques, such as for example, ranking of magnitudes in order, a determination is made as to which of the forty eight values are highest. A fixed, predetermined number of highest values are selected (step 84 of Fig. 4). For example, in accordance with a preferred embodiment of the
20 invention, it has been found useful to determine which of registers R1 to R48 have the twelve highest values.

The twelve highest values selected by arithmetic processing routine 52 are transfers to respective registers 54a to 54i (there are also forty eight of these
25 registers). This is represented in step 86 of Fig. 4. All other registers of registers 54a to 54i are loaded with or retain a value of zero (step 88 of Fig. 4). The steps described above are repeated for the next frame of the source material, and the contents are stored in
30 registers 56a to 56i. At this point, registers 54a to 54i have contents as described above, representative of a

first frame of the source signal, while registers 56a to 56i have contents representative of the second frame. This process is repeated for three additional frames, for a total of five frames, and the contents are successively
5 stored in registers 58a to 58i, registers 60a to 60i, and registers 62a to 62i, until these buffers have contents, as described above, representative of five successive frames of the signal.

When the registers have been filled as described above,
10 an arithmetic averaging routine, represented as 64, will average the contents of all of the registers 54, 56, 58, 60 and 62 along a row of registers in Fig. 3 (step 90 in Fig. 4). The contents are then placed in respective registers 66a to 66i. In other words, register 66a
15 contains the average value of registers 54a, 56a, 58a, 60a and 62a. Register 66b contains the average value of registers 54b, 56b, 58b, 60b and 62b. Similar statements may be made for the remaining ones of the forty eight registers as represented by 66c through 66i.

20 The values stored in registers 66a to 66i are processed by an output processing routine 68. Output processing routine 68 is similar to arithmetic processing routine 52 in that it is designed to determine the highest values of the averages stored in registers 66a to 66i (step 92 in
25 Fig. 4). For example, in accordance with a preferred embodiment of the invention, it has been found useful to determine which of registers 66a through 66i have the highest eight values. These eight values are represented by a binary value of "1" in respective registers 70a to
30 70i. The remaining one of registers 70a to 70i remain or are set to a value of binary "0". The processing by

output processing routine 68 is represented by step 94 in Fig. 4.

It is noted that near silent frames may be represented by a predetermined unusual and unique 48-bit value with
5 eight bits set in an inharmonic pattern extremely unlikely to occur in source material. The value [100000100000100000100000100000100000100000100000000] is exemplary. Given a root frequency of B, this would yield the following musical notes: B - F - Bb - E - A - Eb - Ab
10 - D. This is just one of many possible heuristic values. This value represents two chromatic clusters (D,Eb,E,F and Ab,A,Bb,B) distributed to form four tritones (B/F, Bb/E, A/Eb, Ab/D) and six major seventh (B/Bb, F/E, Bb/A, E/Eb, A/Ab, Eb/D) intervals. The clusters and intervals
15 are dissonant in and of themselves, and it is extremely unlikely that they would occur together in music.

It is also noted that the highest twelve and eight values are represented as floating point numbers, and a tie as
20 to which frequency is of highest amplitude is unlikely. However, if a tie does occur, then either the higher represented frequency of the forty eight frequencies or the lower represented frequency is designated as the higher twelfth or eighth value. It is not critical as to
25 whether it is the higher or lower of the two frequencies that is represented, as long as this is done on a consistent basis. After five frames of the signal have been processed as described above, all registers of Fig. 3 may be cleared (reset to have values of zero) and the
30 entire process is repeated for the next five frames. Alternatively, the values are simply written over during processing of the next five frames.

The result of the processing described above is a representation of five successive frames of signal, wherein no matter what the nature of the signal, a fixed, predetermined number of bits (eight bits in the exemplary
5 embodiment) of the total (forty eight in the exemplary embodiment) are always set to "1". Having a constant number of bits per frame set to a particular binary value has marked advantages. First, there is constant information density provided at the output of the data
10 reduction process.

Another advantage is that the probability of having N bits match in any particular frame, when a processed signal is compared to a reference, can be pre-calculated in a fairly precise way. Yet another advantage is that
15 the number of bits representative of any source material is always directly related to the number of frames (8 in the preferred embodiment), and thus does not need to be calculated by, for example, the sum of bit count lookups.

Thus, in the example described above, an array of 48 bits
20 is produced every five frames (every one tenth of a second). A series of successive arrays may compared to data in a reference library to determine whether known program content is present in a broadcast signal being monitored, as for example in United States Patent No.
25 5,437,050.

Utilizing the Reduced Data - Scoring Algorithm

In accordance with the present invention, a determination as to whether there is a match between source data and data in a reference library may be based on the following
30 analysis. In other words, successive frames of the

reduced data and a reference are compared. The basis for the comparison is explained below.

The probability of having n bits match in any frame is:

$$\frac{\binom{8}{n} \binom{40}{8-n}}{\binom{48}{8}}$$

This allows for the computationally efficient assignment of a non-linear match value per frame, thus yielding an excellent curve fitting metric. The bits matched count of a frame (always in the range of zero to eight) is used to look up a match value, rather than to perform a more computationally intensive match calculation as in United States Patent No. 5,437,050.

Match values are assigned heuristically with deference to and consideration of the probabilities of random data having that many bits match. The raw probabilities, calculated as in equation above, for n out of eight bits matching are { .2038, .39, .284, .097, .017, .0014, .000058, .00000085, .0000000027 }. The sum of these probabilities, if taken to the limit of accuracy, totals 1. A match of six out of eight may be taken as a baseline of a probability of 1.0. The values then become:

| | |
|---|---------------------|
| | 1 - 6,500 |
| | 2 - 4,733 |
| | 3 - 1,616 |
| | 4 - 283 |
| 5 | 5 - 23 |
| | 6 - 1 |
| | 7,8 - better than 1 |

The square roots of the resulting numbers are taken, they are multiplied by 10, and subtracted from 1.000. The values for seven and eight are simply assigned as .999 and 1.000, respectively, so that a value greater than .990 (the value for $n=6$) is produced. However, this is a heuristic construct, as there is insufficient scale for representing how much better a match with seven or eight bits (as compared to a match with six bits) really is.

The resulting table of match values based on the number of bits that are the same, over the range of zero through eight bits, is:

(0, .194, .312, .589, .832, .952, .990, .999, 1.000).

20

The match values may be summed and averaged over the comparison run length of the sample and the reference in the reference library to yield the score.

The score for k frames for any two sets of data A and B may be computed as:

$$\begin{array}{rcl}
 & k-1 & \\
 & \sum & \text{match Value(bits Matched(Ai,Bi))} \\
 & i=0 & \\
 5 \quad \text{Score} & = & \frac{\quad}{k}
 \end{array}$$

For purposes of discussion, as an example, a small sample of reduced data, produced in accordance with the invention, is used. If in one second, there are ten frames of reduced sample data to match to ten frames of reference data, a table of number of matched bits and corresponding match values, in accordance with the criteria set forth above is represented as:

| | <u>Number of Bits Matched</u> | <u>Match values</u> |
|----|-------------------------------|---------------------|
| 15 | 6 | 0.99 |
| | 7 | 0.999 |
| | 5 | 0.952 |
| | 3 | 0.598 |
| | 5 | 0.952 |
| 20 | 2 | 0.312 |
| | 4 | 0.832 |
| | 3 | 0.598 |
| | 6 | 0.99 |
| | 5 | <u>0.952</u> |
| 25 | Total | 8.175 |
| | Score (= average) | 0.8175 |

Depending on the application, the score of 0.8175 may indicate that there is a match between the sample and the reference data. Thus, it has been found that the present invention provides excellent short-term discrimination, with highly separable scores between matches and non-matches. Further, when the phase of the sample signal is shifted with respect to the reference, performance remains excellent, showing that matching is not phase sensitive. Finally, less data must be processed than in prior art systems, and thus computational efficiency is greatly enhanced.

If a determination has been made that a match is unlikely, then it is assumed that the program content is not in the reference library. Program content that is not in the reference library may be evaluated in the manner described in United States Patent No. 5,437,050 to determine which parts of the program content is a "suspect" and to allow a human operator to determine which suspects represent content that should be added to the reference library.

Variations described for the present invention can be realized in any combination desirable for each particular application. Thus particular limitations, and/or embodiment enhancements described herein, which may have particular advantages to the particular application need not be used for all applications. Also, it should be realized that not all limitations need be implemented in methods, systems and/or apparatus including one or more concepts of the present invention.

The present invention can be realized in hardware, software, or a combination of hardware and software. Any kind of computer system, or other apparatus adapted for carrying out the methods and/or functions described herein, is suitable. A typical combination of hardware and software could be a general purpose computer system with a computer program that, when being loaded and executed, controls the computer system such that it carries out the methods described herein. The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods described herein, and which, when loaded in a computer system, is able to carry out these methods. Computer program means or computer program in the present context include any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after conversion to another language, code or notation, and/or reproduction in a different material form.

Thus the invention includes an article of manufacture which comprises a computer usable medium having computer readable program code means embodied therein for causing a function described above. The computer readable program code means in the article of manufacture comprises computer readable program code means for causing a computer to effect the steps of a method of this invention. Similarly, the present invention may be implemented as a computer program product comprising a computer usable medium having computer readable program

code means embodied therein for causing a function described above. The computer readable program code means in the computer program product comprising computer readable program code means for causing a computer to
5 effect one or more functions of this invention. Furthermore, the present invention may be implemented as a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for causing one or more
10 functions of this invention.

It is noted that the foregoing has outlined some of the more pertinent objects and embodiments of the present invention. The concepts of this invention may be used
15 for many applications, as discussed above. Thus, although the description is made for particular arrangements and methods as an examples, the intent and concept of the invention is suitable and applicable to other arrangements and applications. It will be clear to
20 those skilled in the art that other modifications to the disclosed embodiments can be effected without departing from the spirit and scope of the invention. The described embodiments ought to be construed to be merely illustrative of some of the more prominent features and
25 applications of the invention. Other beneficial results can be realized by applying the disclosed invention in a different manner or modifying the invention in ways known to those familiar with the art. Thus, it should be understood that the embodiments has been provided as an
30 example and not as a limitation. The scope of the invention is defined by the appended claims.